

DOCUMENT RESUME

ED 403 289

TM 026 008

AUTHOR Hambleton, Ronald K.
TITLE Setting Standards on Performance Assessments:
Promising New Methods and Technical Issues.
PUB DATE Aug 95
NOTE 12p.; Paper presented at the Annual Meeting of the
American Psychological Association (New York, NY,
August 1995).
PUB TYPE Information Analyses (070) -- Speeches/Conference
Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Cutting Scores; Holistic
Evaluation; *Licensing Examinations (Professions);
*Pass Fail Grading; *Performance Based Assessment;
Profiles; Scores; Simulation; Standards; Statistical
Distributions; *Test Construction; Test Length
IDENTIFIERS Angoff Methods; Compensatory Models; *Conjunctive
Item Response Functions; Contrasting Groups Method;
Policy Capturing Method; *Standard Setting

ABSTRACT

Performance assessments in education and credentialing are becoming popular. At the same time, there do not exist any well established and validated methods for setting standards on performance assessments. This paper describes several of the new standard-setting methods that are emerging for use with performance assessments and considers their strengths and weaknesses. Methods described that are being applied to performance assessments are: (1) contrasting groups; (2) extended Angoff; (3) estimated mean, expected score distribution; (4) paper selection; (5) holistic or booklet; (6) dominant profile; and (7) policy capturing. A special problem is that of compensatory versus conjunctive standard setting. A compensatory standard is one in which any candidate who achieves a defined total score will pass. In conjunctive standard setting, raters set a conjunctive standard by stressing the most important parts of the assessment or making performance on a given item decisive rather than relying on overall score. A simulation study that compared the reliability and validity of these two approaches found surprising results for the conjunctive standard that suggest that increasing numbers of candidates will fail as the assessment length increases, and validity will actually decline. More research is needed to find better ways to set standards on performance assessments, although substantial evidence shows that defensible standards can be set for achievement and credentialing performance examinations. (Contains 4 figures and 12 references.) (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

- ☒ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Setting Standards on Performance Assessments: Promising New Methods and Technical Issues

Ronald K. Hambleton
University of Massachusetts at Amherst

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY

RONALD K. HAMBLETON

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

Abstract

Performance assessments in education and credentialing are becoming popular. At the same time, there do not exist any well established and validated methods for setting standards on performance assessments. The purposes of this paper are (1) to describe several of the new standard-setting methods which are emerging for use with performance assessments and consider their strengths and weaknesses, and (2) to consider a special problem, that of compensatory versus conjunctive standard-setting methods. The main conclusions are that there is plenty of room for new ideas, creativity, and research in standard-setting methodology, and more effort is needed to document and validate standards for intended uses.

Setting Standards on Performance Assessments:
Promising New Methods and Technical Issues^{1,2}

Ronald K. Hambleton
University of Massachusetts at Amherst

1. Introductory Remarks

Whenever important individual decisions are made with examinations, there will be challenges to how the performance standards were set and who set them. Performance standards are the "Achilles heal" of credentialing exams. The fact that different standard-setting methods, in general, lead to different results (i.e. standards) is just one of many troublesome features of performance standards for credentialing agencies to explain (Zieky, 1995).

Fortunately, there have been many successful efforts to set standards on multiple-choice credentialing exams. The defense of performance standards for a particular credentialing examination is based on (1) the credibility of the process used to set them (e.g. selection of appropriate panelists, excellent panelist training, and a well-planned and systematic process which provides ample opportunity for discussion and deliberations among the panelists), (2) the reliability of the performance standards (that is, there needs to be agreement among panelists about the performance standards), and (3) the reasonableness of the performance standards (i.e., the passing rate is not too far out of line with expectations about the quality of persons entering the profession). Current standard-setting methods can produce defensible results. Of course it is also true that not all credentialing agencies commit the resources and time to set standards in a defensible manner.

Now, there is a new challenge to performance standard-setting methods: performance assessments. In both educational testing and credentialing exams, there has been an emergence of performance assessments. Kentucky has moved to a total performance assessment system for student and school accountability. Nearly all other states are using some form of performance assessment in student accountability. Most of my own standard setting research in the last two years has been with Dick Jaeger and Barbara Plake in developing new methods for setting standards on the performance-based assessments of teachers (see, for example, Jaeger, Hambleton, & Plake, 1995). I have also been working on performance assessments such as "standardized patient assessments" in the medical examination area with the Education Commission for Foreign Medical Graduates (Vu & Barrows, 1991).

¹Laboratory of Psychometric and Evaluative Research Report No. 278. Amherst, MA: University of Massachusetts, School of Education.

²Paper presented at the meeting of APA, New York, August, 1995.

Performance assessments are often associated with complex scoring rubrics, multidimensionality in the response data, and low score reliability at the task or exercise level. These features create special problems for standard-setting methods. In addition, several of the popular standard-setting methods that Mike Zieky reviewed in his paper (Zieky, 1995), such as the Nedelsky and Angoff methods, are not applicable with performance assessments. The challenge for psychometricians is to develop new and defensible standard-setting methods to meet the current characteristics of performance assessments.

In the remainder of this presentation, I plan (1) to describe several of the new methods which are emerging for use with performance assessments, and (2) to consider a special problem, that of compensatory versus conjunctive standard-setting methods. From our experience, panelists often want to be conjunctive in setting standards - they say things like, if a candidate fails this task, then he/she deserves to fail (Hambleton & Plake, 1995). We regularly hear this type of statement from panelists in our standard-setting work. But, when exercise scores have low levels of reliability (and this situation is common), conjunctive standard-setting policies are flawed and this point will be addressed in more detail later in the presentation.

2. Standard-Setting Methods for Performance Assessments

My plan in this section is to provide a list of several methods which might apply to performance assessments and to provide brief descriptive and evaluative comments. With several methods, follow-up references are provided.

Contrasting Groups. Some researchers have recommended this method but this method would rarely be feasible with credentialing exams. It simply isn't possible, typically, to identify candidates independently of the credentialing examination who are masters and non-masters. And even if the mastery status of candidates could be identified, it would be hard to argue that the candidate samples of masters and non-masters are representative of the populations of masters and non-masters, respectively. Without representative sampling, any resulting performance standard would be sample dependent and of little value. The method may be more promising with performance assessments in school settings where teachers can be used to determine mastery status of students independent of the exam. Additional concerns about the contrasting groups method are presented by Kane (1994).

One promising exception is the work of Clauser and Clyman (1994) who identified passing and failing performance based on a panel's holistic review of candidate exam performance. Then the score distributions of these two groups of candidates were used in setting performance standards. This method is limited, however, by the use of an internal criterion.

Extended Angoff. This method appears to have some promise with performance assessments (Hambleton & Plake, 1995). Panelists are required to determine the number of score points on exercises that would be obtained by borderline candidates (that is, candidates just skilled and knowledgeable enough to deserve credentialing). Panelists can even set weights for exercises in arriving at composite scores. Exercises judged as more important can

be assigned higher weights in the scoring. Our research suggests that the method is popular and can lead to acceptable performance standards. It is however a compensatory method, a feature which not all panelists approve of. This means that candidates can compensate for low performance on some exercises by achieving higher scores on other exercises and still meet the performance standard for the exam.

Panelists need to be intimately familiar with the scoring protocols for the exercises. In some of our recent work, two to three days was spent familiarizing panelists with the scoring protocols. The standard-setting process itself may take another day or two (Hambleton & Plake, 1995).

Estimated Mean, Expected Score Distribution. This method has similarities to the extended Angoff method. Here, panelists are required to not only determine the minimum number of score points for borderline candidates, but they also estimate the distribution of scores of the borderline candidates. The method is being tried by the National Assessment Governing Board (NAGB) in their work to set performance standards on the National Assessment of Educational Progress (see, for example, Cooper-Loomis & Bourque, 1996).

The advantage of this method, in principle, is that additional relevant information about the performance of borderline candidates is extracted from panelists. Panelists, who expect the standard deviation of the score distribution for borderline candidates to be low, are expressing considerable confidence in the placement of the standard. Higher standard deviations correspond to less confidence on the part of panelists.

Paper Selection. Here, panelists are instructed to identify papers/work/projects etc. from the assessment which they believe are associated with borderline candidates. After some discussion, revised selections can be made. The average score associated with the papers identified as borderline is chosen as the performance standard. This method is being used by NAGB, by some state departments of education, by the National Board of Medical Examiners, and the Educational Commission for Foreign Medical Graduates, on an experimental basis. The big advantage, from my perspective, is that panelists are required to look at the work of candidates. The big disadvantage is that the method can be time consuming and awkward to implement in practice. For example, when the candidates' work involves video-tapes, reports, projects, etc., sorting through candidate work for examples of borderline work can be very tedious, if not totally impractical. Still, the paper selection method is deserving of considerably more research and development. This method is being field-tested currently in a performance standard setting project funded by the National Science Foundation (see, Hambleton, Jaeger, Plake, and Mills, 1996).

Holistic or Booklet. This method has some similarities with the paper selection method. This is a new method suggested by the National Academy of Education in its review of the standard-setting work of NAGB and the American College Testing (ACT) (Shepard, Glaser, Linn, & Bohrnstedt, 1993). Basically, panelists look at the complete work of candidates and decide which ones are borderline (or masters and non-masters). It has been suggested as an alternative to the Angoff method with multiple-choice items too. The criticism is that with the Angoff method, focused at the item level, the overall picture of a

candidate's performance is lost. NAGB and ACT have been field-testing this method with National Assessment of Educational Progress (NAEP) data. It remains to be determined how well the method will work in practice. Certainly the focus on candidate work is admirable.

Dominant Profile. This method is a direct assault on the standard-setting problem. A panel, after becoming familiar with the purpose of the assessment and the scoring scheme, tries to formulate a standard-setting policy. A standard-setting policy might be something like:

A candidate passes the exam if he/she has (1) an overall score of 18 on the seven exercise assessment, (2) scores of at least "3" (out of 4) on exercises B and C, the two exercises judged to be most important, and (3) no scores of "1" on the exercises which indicate disappointing/unsafe performance.

The method may begin with a consideration of which profiles of scores over the exercises are worthy of credentialing. Over a series of iterations, the panel tries to arrive at a consensus policy or set of rules for passing and failing candidates. No limits or restrictions are placed on the final result. It may be compensatory, conjunctive, or some combination. The big advantages are that the method is direct and involves extensive discussions among panelists. From our experiences, panelists find the discussions very helpful. They have appeared suspicious of methods which they cannot completely control. The big disadvantage is that a single policy for making pass-fail decisions may not emerge from the panel. For example, suppose the panel is fundamentally divided on the desirability of a conjunctive component in the policy (such as components 2 and 3 in the example above). And, unlike the performance standards set with other methods, it may not be possible to average policies to arrive at a group consensus. This method has been studied by Plake, Hambleton, and Jaeger (in press). Another disadvantage is that a conjunctive policy may result which is based on unreliable exercise scores (see, Hambleton & Slater, in press).

Policy Capturing. This method involves panelists considering hypothetical score profiles and judging their level of competence. Then, a mathematical model (a linear regression line) is fit to a panelist's ratings to determine his/her standard-setting policy. A group policy (or standard-setting procedure) can be obtained by a weighted average of the individual panelist's ratings. Successive iterations are used to make panelists more consistent in their ratings, and to move the group of panelists toward a consensus policy for making pass-fail decisions. A big advantage is that a result is guaranteed. Potential disadvantages are that it may be difficult to find statistical models to fit individual panelist's ratings of the score profiles, and the mathematical manipulations of the data make it difficult to explain to panelists what exactly is done with their ratings. Some researchers believe that panelists ought to completely understand the process used in arriving at the standard. This method has been under development by Richard Jaeger for several years and the results, to date, are very encouraging (see, for example, Jaeger, 1995). More research is currently underway (see, Jaeger, Hambleton, & Plake, 1995).

3. Compensatory versus Conjunctive Standards

Many performance assessments consist of a set of tasks or exercises to be completed by candidates. For example, in the certification process for foreign medical graduates, candidates work their way through 10 medical cases involving standardized patients. Through these cases, the medical competencies of candidates can be assessed. In the case of the National Board for Professional Teaching Standards, candidates may be presented with 6 to 8 exercises. Some exercises are completed in the candidates' schools and may involve such activities as the taping of lessons and the preparation of classroom logs. Other exercises such as preparing a lesson or analyzing a lesson are completed at the assessment center.

Consider a performance assessment or exam with (say) eight exercises, each scored on a four point scale. How might performance standards be set? The extended Angoff method would result in setting a standard on each exercise and then summing those standards to arrive at a standard for the total assessment. Any candidate who achieves that total score would be passed and this is known as a compensatory standard. Any combination of performance scores across the exercises is certifiable if the total candidate score exceeds the performance standard. But panelists often want to do something different: they will say that "I want to fail any candidate who does not achieve a score of at least 3 out of 4 possible points on exercise 2." This is known as a conjunctive standard. Panelists will say things like, "this exercise 2 is far and away the most important and if candidates don't score well, they do not deserve certification." Or they might say, "any score of 1 on a four point scale is horrible medical performance-these candidates deserve to fail, regardless of how they may have performed on other exercises." In some of our recent work, for example, panelists have expressed a definite preference for a conjunctive standard (Hambleton & Plake, 1995).

I plan to report on a small simulation study (Hambleton & Slater, in press) to compare the reliability and validity of compensatory and conjunctive standard-setting methods. Two additional variables which impact on the findings are the number of exercises in the assessment and the correlation between pairs of exercises. Both of these variables were manipulated in the study. Reliability was assessed as the consistency of pass-fail decisions across parallel administrations. Validity was assessed as the agreement between pass-fail decisions resulting from the assessment itself and the true state of candidate knowledge (which, of course, can be known in a simulation study).

Each exercise was scored on a four-point scale (1 to 4); the passing score on each exercise was set at 2.5. With the compensatory standard, candidates were required then to obtain a score of 50% or greater across the exercises (for example, with 8 exercises, a score of 20 points-- 8×2.5 -- was needed to pass). With the conjunctive standard, candidates were required to obtain a total score of 50% or greater across the exercises and candidates were not permitted to have any scores of 1 on exercises. This is the conjunctive component of the standard and reflects a common preference of panelists setting standards on certification exams (see, for example, Hambleton & Plake, 1995; Plake, Hambleton, & Jaeger, in press; Jaeger, Hambleton, & Plake, 1995).

Main results from this study are contained in Figures 2, 3, 6, and 7 from the paper by Hambleton and Slater (in press). The results for the compensatory standard behave in the expected ways: The more exercises the better, reliabilities and validities go up with the number of exercises, and more highly correlated exercises lead to better results for any particular number of exercises. The results for the conjunctive standard behave in unusual ways: reliabilities increase with the number of exercises but in a surprising way. For substantial numbers of exercises, the reliabilities are higher for the lower intercorrelations. This is due to the ever increased failure rate associated with the conjunctive standard. For long assessments, nearly all candidates will fail, and this finding is repeated in a parallel administration. And, as for validity (see, Figure 7), validity actually goes down with an increase in the number of exercises! This result is surprising, but upon reflection, makes sense. With the unreliable scores at the exercise level, even the best candidates will occasionally obtain a score of 1 (due to measurement errors in the assessment), and with the conjunctive standard in place in the study, these candidates will fail the assessment. If the assessment is made quite long, then nearly all candidates will fail. The resulting decisions will be inconsistent with their true state of knowledge, and validity indices will go down. Ironically, the longer the assessment, the worse the validity results!

I am not suggesting that all conjunctive standards will produce such results, but this popular one, that has arisen in several recent projects, certainly would produce disappointing results in practice. In one recent study with an actual credentialing exam, with such a component in place, only 4% of candidates would have been certified. If exercise reliability were higher, or if the conjunctive component were less common (for example, fail a candidate with a score of 1 on a 10 point exercise with a passing score at 6), then less problems can be expected. The overall problem seems due to the fact that when panelists are setting standards they fail to consider the role of measurement errors in candidate performance. In our current work, we are advising against the use of conjunctive components but more research on this important point is needed. For one, how could you train panelists to consider measurement errors in their ratings? They already seem overloaded with relevant information for implementing a standard-setting process with several of the promising methods described above.

4. Concluding Remarks

Twenty five years of research and development has improved the approaches to setting standards and I believe that there is now substantial evidence to show that defensible standards can be set for achievement and credentialing exams. At the same time, performance assessments present new challenges to measurement specialists: new methods are in their infancy, and need to be fully evaluated; there is plenty of room for creativity in the development of new methods (possibly drawing from other disciplines, policy capturing came from I/O psychology, for example). At the same time, at least some methods may need to be shortened (NAEP and NBPTS are spending six days per exam) but in my experience in the medical profession, anything more than two days can't even be considered. Finally, validation initiatives are central to the defensibility of any set of standards. Compilation of substantial amounts of evidence to support the intended uses of the standards are needed.

Having a carefully selected and well-trained panel of judges who deliberate carefully may not be sufficient to defend a standard.

The main points of my conclusion:

1. There is great need to further research promising new methods while at the same time, there is considerable room for new ideas and creativity. We are not close to closure on the best methods for setting standards on performance assessments.
2. More commitment is needed to document the implementation of a standard-setting method and more initiatives are needed to validate standards for particular uses.

References

- Clauser, B. E., & Clyman, S. G. (1994). A contrasting-groups approach to standard setting for performance assessments of clinical skills. Academic Medicine, 69(10), S42-S44.
- Cooper-Loomis, S., & Bourque, M. L. (1996, April). Psychometric considerations of items for achievement levels setting. Paper presented at the meeting of NCME, New York.
- Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (1996). Methods for setting standards on performance assessments: review of issues and methods (Laboratory of Psychometric and Evaluative Research Report No. 284). Amherst, MA: University of Massachusetts, School of Education.
- Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. Applied Measurement in Education, 8, 41-56.
- Hambleton, R. K., & Slater, S. C. (in press). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. Applied Measurement in Education.
- Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. Applied Measurement in Education, 8, 15-40.
- Jaeger, R. M., Hambleton, R. K., & Plake, B. S. (1995, April). Eliciting configural performance standards through a sequenced application of complementary methods. Paper presented at the meetings of AERA and NCME, San Francisco.
- Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425-462.
- Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (in press). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. Educational and Psychological Measurement.
- Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). Setting performance standards for achievement tests. Stanford, CA: National Academy of Education.
- Vu, N. V., & Barrows, H. S. (1991). Use of standardized patients in clinical assessments: Recent developments and measurement findings. Educational Researcher, 23, 23-30.
- Zieky, M. (1995, August). Aspects of standard-setting methodologies. Paper presented at the meeting of APA, New York.

Figure 2. Decision consistency with a total score compensatory standard-setting policy

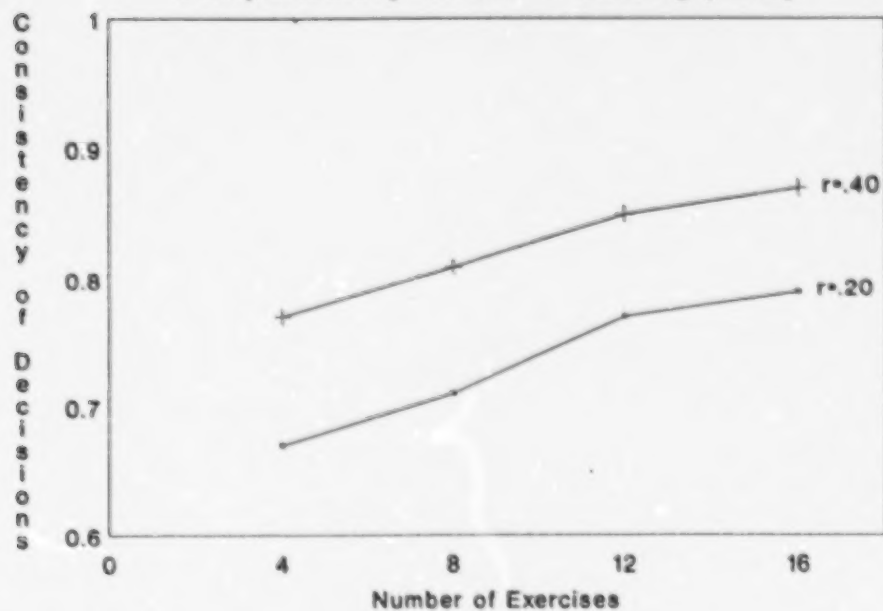


Figure 3. Decision accuracy with a total score compensatory standard-setting policy

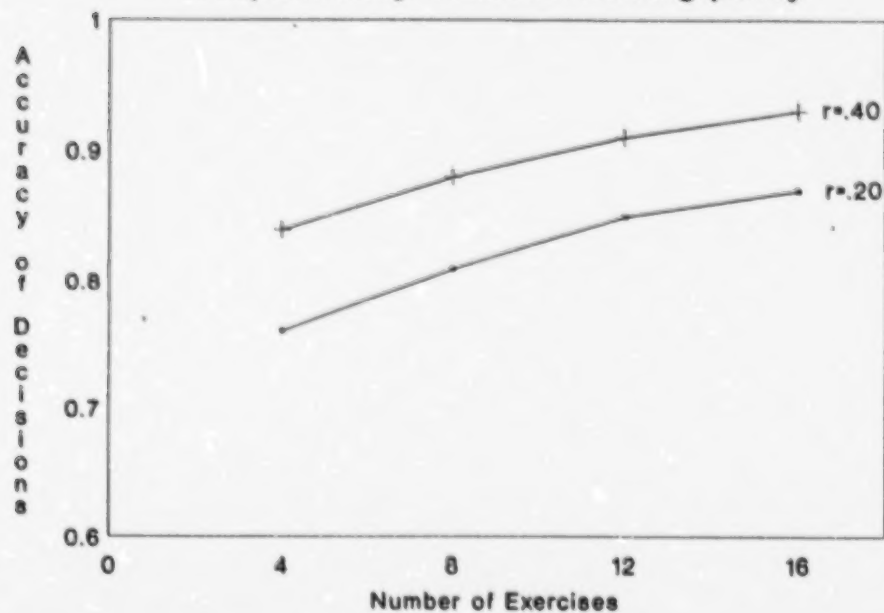


Figure 6. Decision consistency with a total score compensatory standard-setting policy with an additional conjunctive component (no scores of '1')

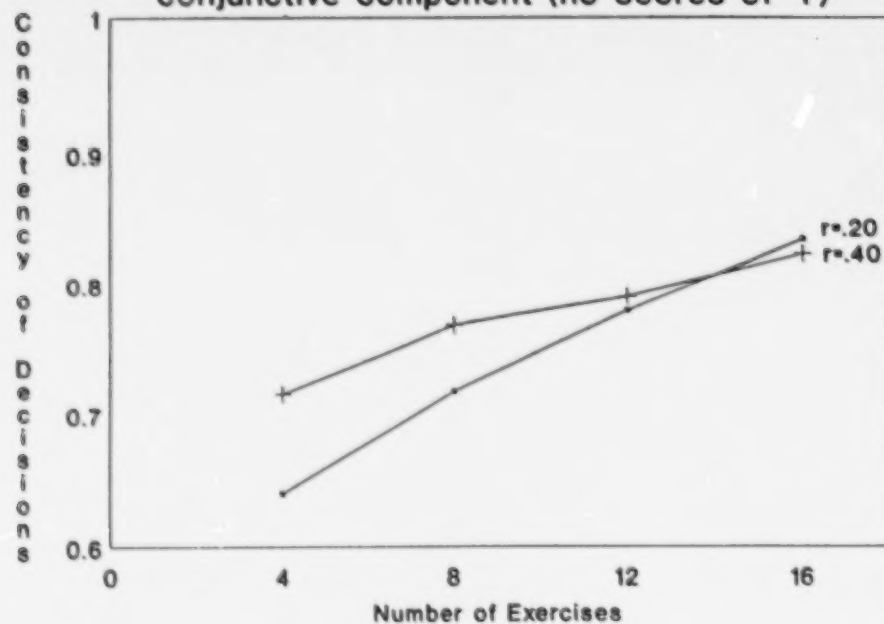


Figure 7. Decision accuracy with a total score compensatory standard-setting policy with an additional conjunctive component (no scores of '1')

